Materials and Methods

Virus selection and database construction

We selected single-stranded RNA (ssRNA) viruses as our focal virus group because they are a primary pathogen group responsible for emerging human disease (*2*, *4*). Our dataset spanned 12 taxonomic groups of positive and negative sense ssRNA viruses that contained human pathogenic species (*20*) and covered included 80% of ssRNA virus families that contain human-infecting species, data from ref. (*4*). The only human-infecting ssRNA viral families we excluded were the *Retroviridae*, which might not be representative of other ssRNA viruses due to their distinctive life cycle involving integration into the genome of host cells; *Orthomyxoviridae*, which can have different host origins in each genomic segment and would have required a different modelling approach; and the *Bornaviridae*, which contains only 8 species. Eleven of the 12 groups we studied are classified as viral families by the International Committee for Viral Taxonomy (ICTV). The remaining group, the order *Bunyavirales* (formerly family *Bunyaviridae*), contained viruses from 7 families (*Feraviridae*, *Hantaviridae*, *Jonviridae*, *Nairoviridae*, *Peribunyaviridae*, *Phenuiviridae* and *Tospoviridae*). All analyses were qualitatively equivalent when analyzing the Bunyaviruses as a single order or as families. Most records were identified at the viral species level; however, when available, phylogenetically divergent and epidemiologically district strains (e.g., maintained by different species in different geographic regions) were analyzed independently.

Using authoritative texts on viral families and the primary literature, we recorded the primary reservoir host taxon, evidence of arthropod involvement in transmission and the taxonomic group of arthropod vectors for each virus, where known (Appendix S1). We defined reservoirs as the host group currently accepted to be responsible for the long-term perpetuation of each virus (also referred to as natural host or maintenance host). To maximize host-specific signals within the viral genomes, viruses suspected to have multiple reservoir taxa that spanned >1 host class analysis were excluded from training. We used a single representative genome from each virus species/strain to avoid inflating model accuracy by including the same viral taxa in model training and validation. Single genome prediction is also more realistic for application to newly discovered viruses that lack multiple reference sequences.

Reservoir host categories were operationally defined to accommodate a trade-off between the granularity of predictions and the number of representatives per group, while maximizing the utility of predictions for disease mitigation and surveillance. Specifically, we aimed to subdivide birds and mammals, the two main reservoir groups associated with emerging zoonotic disease (*2*), to the maximum extent possible to increase the resolution of our predictions without creating undersized or excessively unbalanced host classes. Birds and mammals were therefore progressively split to the next lower taxonomic level where this maintained at least two sub-classes with sample sizes ≥15. Bats (order Chiroptera) were split into Pteropodiformes (families Pteropodidae, Rhinolophidae, Hipposideridae, Megadermatidae, and Rhinopomatidae, here abbreviated "Pterobats") and 'Vespertilioniformes' (remaining microbat families, here abbreviated "Vespbats") (*21*). Rodents (order Rodentia), are taxonomically diverse hosts of many viruses, but could not be split because available viruses were disproportionately found in only one subgroup (suborder Myomorpha, mouse-like rodents). Similarly, splitting the carnivores (order Carnivora) would have retained only a single host group above our sample size threshold (suborder Caniformia), so carnivores were analyzed at the order level. Birds were split at the superorder level into 2 classes, Neoaves (most modern birds) and Galloanserae (fowl). Our

final analysis used eleven categories of reservoir groups, including insect-specific and plant-specific viruses. While insect-specific and plant-specific viruses are unlikely to cause illness in vertebrates, they are useful categories to include for surveillance or metagenomic applications. Specifically, our models could help determine whether samples from vertebrates may have an insect or plant origin (e.g., dietary viruses or contamination) rather than being potential pathogens of the vertebrate and whether viruses discovered in arthropods are likely to be arthropod-specific or arthropod-borne viruses of vertebrates. Rare reservoir host groups, defined as those with <15 virus representatives (Cetacea [N = 2], Diprotodontia [N = 2], Erinaceidae [N = 1], Lagomorphs [N = 6], Macropods [N = 1], Perissodactyls [N = 4], Reptiles [N = 8], Scandentia [N = 3] and Sorciomorpha [N = 1]) were excluded from model training, but were retained to assess model performance on viruses from un-represented host groups (Figure S18). Viral families contained between 3 and 62 virus taxa (mean = 36.4) with known reservoirs within one of eleven groups. The average virus group contained 7.18 reservoir host classes (range = 2-10). We used a similar strategy to assign arthropod vectors to 4 main groups associated with viral transmission to vertebrates (midges, mosquitoes, sandflies and ticks), but lowered our threshold to 8 viruses to accommodate the lower sample sizes available when analyzing only vector-borne viruses. Four of twelve viral groups contained viruses transmitted by one of four main classes of blood-feeding arthropod vectors, with an average of 3 vector taxa per viral group (Figure 1A).

Quantification of viral genomic traits

For each virus species or strain, all of the coding sequences from a single representative genome were downloaded from Genbank, using NCBI reference sequences (RefSeqs) when available. Coding sequences covered complete genomes for most viruses, but large fragments and single segments (for segmented virus) were also included (N = 11). When reference sequences were unavailable, we selected a random sequence, excluding sequences labelled with the terms "vaccine", "construct", or "recombinant." Nucleotide [A, C, G, T, N] counts were summed across all the coding sequences and represented as a proportion of the total coding sequence length of the selected species isolate; all ambiguous bases were recorded as N. Dinucleotide bias (relative frequencies) for each of the 16 possible dinucleotides was calculated across all coding sequences using the following formula:

$$\frac{\left(\frac{N_{XY}}{DN_{tot}}\right)}{\left(\frac{N_X}{N_{tot}} \times \frac{N_Y}{N_{tot}}\right)},$$

where N_X and N_Y are the total counts of nucleotides X and Y, respectively, N_{tot} is the total number of nucleotides, N_{XY} is the total count of the dinucleotide XY, and DN_{tot} is the total number of dinucleotides across all coding sequences of the selected isolate. Because the effects of dinucleotide bias on viral fitness are reported to be strongest at the bridge between adjacent codons (*22*) and because dinucleotide biases across codon positions were poorly correlated in our dataset (Figure S1), we further calculated dinucleotide bias at "bridge" and "non-bridge" codon positions.

Codon pair bias was measured as the codon pair score (CPS) for each of the 4,096 (64 x 64) possible codon-codon pairs. For completeness, stop codons were included to record rare cases of read through stop codons. We calculated the CPS of a given codon pair independently of codon and amino acid biases, following (*23*):

$$CPS = \ln\left(\frac{AB}{\frac{A \times B}{X \times Y} \times XY}\right),$$

where A and B represent the observed counts of codons A and B, respectively, AB represents the observed count of codon pair AB, X and Y represent the observed counts of the corresponding amino acids X and Y, respectively, and XY represents the observed count of amino acid pair AB across all coding sequences of the selected species isolate. This CPS score determines if a given codon pair is over-represented (+) or under-represented (-). To avoid null values in cases where an amino acid pair was not observed, all codon pairs that encode for that amino acid were given the average CPS score for the species isolate. When codon pairs were not observed, but the encoding amino acid pair was observed, the codon pair was given a CPS score of -9999 to indicate extreme under-representation.

Codon bias was calculated for each of the 64 codons by dividing the total count of each codon by the total count of all codons that encode for the corresponding amino acid or stop codon across all coding sequences of the selected isolate. Amino acid bias for each of the 21 amino acids (stop is considered as an amino acid here) was calculated by dividing the total count of each amino acid by the total number of amino acids in the isolate. The total number of genomic traits considered was therefore 4229 (CPS = 4096, dinucleotide biases = 48, codon biases = 64, amino acid biases = 21).

Association tests of viral genomic traits with reservoir hosts, vector types and viral taxonomic group

To test broad-scale associations of viral genomic traits with host group, vector group and viral group, we fit macroevolutionary (trait) models to dendrograms estimated by applying unsupervised hierarchical clustering to all 4229 viral genomic traits. Clustering used the Ward (ward.D2) method (*24*), but results were qualitatively similar with other clustering methods. Null distributions for trait models were generated by randomly shuffling either virus group, host group, or vector group along the tips of dendrograms. We compared the fit of 500 models with random trait permutations to the true virus group/reservoir host/vector associations using the *fitDiscrete* function in the *geiger* package of R using an equal rates model without branch length transformations (*25*). Model comparisons used the difference between the average Akaike Information Criterion (AIC) value from 500 trait-randomized trees and the AIC value of models fit to true associations.

We tested effects of reservoir host and arthropod vector associations on viral genomic biases using generalized linear mixed models (GLMMs). Two sets of models were compared using AIC. These focused on viral dinucleotide biases as a computationally manageable subset of all genomic features. First, we compared GLMMs with a random effect of viral taxonomic group and a fixed effect of reservoir/vector to a random effect only model to test effects of host associations on dinucleotide biases while controlling for the effects of virus evolutionary history. Second, to test whether effects of reservoir associations arose from convergence of biases in viruses from different taxonomic groups that persist in the same reservoirs, we compared models with and without a fixed effect of reservoir/vector using a nested random effect of reservoir/vector within viral taxonomic group. We used the Benjamini–Hochberg procedure with a false discovery rate of 0.1 to indicate the significance of *p*-values after correcting for multiple testing (*26*). Figure legends (Figures S2-S7) contain the threshold *p*-values that were considered significant for each set of tests.

Phylogenetic neighborhoods of reservoir hosts, vector types, and arthropod-borne transmission

Clustering of reservoir host and vector taxa on viral phylogenetic trees implies that the reservoir host and/or arthropod vector associations of closely related viral species or strains may inform those of viruses with unknown reservoir hosts or vectors. Incorporating viral phylogenetic information could therefore improve the accuracy of machine learning analyses that use only genomic biases to predict reservoir hosts or vectors. However, since our analysis included highly divergent RNA viruses, homologous genes do not exist to align and build a single phylogenetic tree describing evolutionary relationships. We therefore designed a flexible routine that searched for regions of high sequence similarity between pairs of viruses and constructed reservoir host, arthropod-borne transmission and arthropod vector phylogenetic neighborhoods (PN) comprising the most closely related viruses. Customized local databases of virus sequences with known ecological associations were built using *makeblastdb* in the command line version of BLAST (*27*). For non-segmented viruses, we included complete genomes. For segmented viruses (Arenaviruses and Bunyaviruses), we used only S segments to avoid the possibility of comparing non-homologous segments. We next used *blastn* (default settings except: max_hsps = 1, reward = 2, word_size = 8, gap_open = 2, gapextend = 2) to find the top hits for each virus (excluding hits to the query virus) based on e-values. When many hits exceeded our quality threshold of e-value < 1e-3, viruses were secondarily ranked by bit-score. The top 5 hits were considered the PN of each virus. We calculated the support (*S*) for each reservoir host/arthropod-borne transmission status/vector class (*j*) as a function of the number and pairwise identity of hits to the focal virus:

$$S_j = \sum_j \frac{P_i}{\sum P},$$

where P_i is the pairwise genetic identity between the focal virus and hit *i*. Viruses that had no hits with e-values < 1e-3 were assigned equal *S* across all classes. These analyses were performed using the *ape* and *seqinR* packages of R, *Biopython* package in Python (*28*, *29*). Exploratory analyses using more evolutionarily conserved protein sequences (*blastp*) predicted reservoir hosts less accurately, presumably reflecting the lower resolution of the PN.

The accuracy of PN algorithms without machine learning was calculated by building reference training and validation databases comprised of viruses with known reservoir host or vector associations. We sampled datasets to match the splitting process used in the machine learning analyses to ensure comparability (see below). Specifically, for reservoir host predictions, we randomly sampled 70% of the viruses from each reservoir host group as the reference database (N = 311). Of the remaining viruses, 50% (N = 61) from each host group were discarded to mimic the dataset used for model optimization and 50% were used as a 'holdout' validation set to record model performance (Figure S8). The same splitting regime was used for predicting whether viruses were arthropod-borne. For vector taxa prediction, we randomly under-sampled the majority class (mosquito viruses). Further, to retain a sufficient validation dataset to quantify per class accuracies, we split the data as 60% training, 12% discarded to mimic optimization and 28% retained for validation (i.e. a 30/70 optimization/validation split of the 40% withheld from training). PN-based predictions for test sets were generated by blasting the validation viruses against the reference databases. Among the top 5 blast hits, we recorded the predicted reservoir host group, arthropod-borne transmission status, or vector type with the highest value of S_j and considered this as the PN-based prediction. By comparing the PN-based prediction to the true host/vector associations of each virus, we calculated accuracies as the proportion of correct PN predictions. Uncertainty was estimated by repeating this procedure over 50 splits of reference and validation datasets.

Machine learning analyses

Virus selection and algorithm selection

We used supervised learning approaches to generate functions mapping features extracted from viral genomes to (*i*) reservoir host taxa (11 classes), (*ii*) arthropod-borne transmission (binary) and (*iii*) the taxa of arthropod vectors (4 classes). We first identified promising algorithms by comparing the baseline performance of several parametric and non-parametric methods in predicting the same host or vector class label from the full set of 4229 genomic input features (Figure S9). Specifically, we compared logistic regression (LR), stochastic gradient descent (SGD), naïve Bayes (NB), support vector machines (SVM), k nearest neighbors (KNN), and the tree ensemble methods Extreme Gradient Boosting (XGB), random forests (RF) and gradient boosting machines (GBM). Models were trained using 10-fold cross-validation with random stratified splits to preserve the percentage of samples for each class. Algorithm selection used the following Python libraries: *numpy* and *Pandas* for matrix and dataframe handling; *imblearn*, *scikit-learn*, *xgboost* and *h2o* (*30*) for machine learning; and *matplotlib* for plotting. Remaining machine learning analyses used the *h2o* library in R (*31*, *32*).

Genomic feature selection

As expected from unsupervised clustering (Figure 1), feature importance rankings from algorithm comparisons showed that most of the 4229 genomic biases weakly informed reservoir host and vector associations. To improve the computational efficiency of machine learning algorithms, we conducted a comprehensive search using GBM to select smaller subsets of the most informative features. To ensure that features selected were not tuned to a single training set, we quantified the average relative feature importance in 50 random class-stratified 70% training sets, using 5-fold cross-validation of each training set.

Algorithm optimization and validation

Datasets of viruses with known classes were split into training, optimization and validation sets (Figure S8). Random stratified splitting was used to preserve the percentage of samples for each class. Optimization sets were used exclusively for tuning models during learning and did not contribute to performance estimation. Validation sets were used exclusively to quantify predictive accuracy in 'holdout' data and were not used during training or optimization. For reservoir and arthropod-borne transmission prediction models, 70% of each class was used for training, 15% for optimization, and 15% for validation. For vector taxa, as imbalance between classes impacted performance we under-sampled the majority class (mosquitoes, N = 59) at random (N = 20) to balance class sizes prior to assigning viruses to training/validation sets and used 60% of the dataset for training, 12% for optimization, and the remaining 28% for validation. We optimized each model by grid-searching over a wide range of parameter settings (learning rate, max_depth, sample_rate, col_sample_rate, ntrees, min_rows) to find the combination that maximized accuracy in the optimization set. Up to 500 combinations of parameters were evaluated using a *RandomDiscrete* search. We assessed model performance of each optimized model on the corresponding holdout validation set.

Incorporation of phylogenetic neighborhoods into machine learning analyses

The PN prediction algorithm was built into a pipeline feeding new features into our machine learning training set using local BLAST databases that matched each training set. In contrast to the PN-only based prediction before machine learning (see above), these models were not

restricted to use only the majority prediction within the PN. Instead, PNs were represented as the relative support for each host class within the PN (S_j). For example, PNs for our vector taxa prediction models comprised 4 features describing relative support for midges, mosquitoes, sandflies, and ticks, summing to 1.

Study-wide accuracies

Averaging outputs from multiple models (i.e., "bagging") allowed us to make predictions of the reservoir/arthropod-borne status/vector association of each virus in the study that incorporated uncertainty arising from variation among training sets. Because this approach required each virus to be included in many validation sets (each trained and optimized on different data subsets), we increased the number of models to achieve a minimum of 50 predictions for each virus for each prediction type. This required 550, 600 and 250 class-stratified splits of the data for reservoir host prediction models, arthropod-borne prediction models and vector class prediction models and yielded a median of 82, 130 and 90 observations per virus, respectively. We excluded underperforming models by averaging predictions from the upper 25% of models (based on the overall accuracy) that included the focal virus *i* in the validation set. This approach meant that correct prediction of the focal virus might contribute to a model being included in the top 25%; however, this effect was negligible given that the variation in accuracy observed among validation sets far exceeded the variation that could be attributed to a single focal virus. Consequently, we selected for generally high-performing models, not models that favored the focal virus.

We defined the bagged prediction strength (BPS) for each candidate host as the average predicted probability of that host class across the selected set of models. The highest-ranked host according to BPS was considered the primary bagged prediction. By comparing primary bagged predictions to the recorded host associations of each virus, we were able to infer study-wide per-class accuracies (Figure 2C,G) and virus group-level accuracies (Figure S13H-J). To further quantify model performance, we also recorded the rank of the true host among all possible hosts after bagging predictions across models.

Prediction of unknown reservoir hosts and arthropod vectors of orphan viruses

We applied trained GBMs to predict the unknown reservoir hosts, arthropod-borne transmission status or arthropod vectors of "orphan" RNA viruses using the *h2o.predict* function in the *h2o* package of R (*31*). Predictions described the relative support for each candidate host/arthropod-borne/vector class as BPS. As above, predictions were bagged from 25% of models with the highest validation set accuracies. Given the higher accuracy of the GBMs that combined genomic features and PN for all prediction types (i.e., reservoir, arthropod-borne and vector type), we focused on the projections of those models for most viral groups. Exceptions were reservoir host predictions for orphan Filoviruses and Togaviruses, where we favored the predictions of the GBM using only the genomic features since these models had slightly higher accuracy for those viral groups (Figure S13H).

Temporal evaluation of changes in viral genomic features and model predictions in a new host environment: the 2014-2016 Zaire ebolavirus epidemic

We expected that because host-associated signals in viral RNA involve many substitutions spread across the genome, they were unlikely to be altered by short-term evolution in novel hosts (*7*). To confirm this expectation, we analyzed previously published Zaire ebolavirus (ZE)

genomes collected during the West African epidemic, where a presumed bat-to-human cross-species infection sustained human-to-human transmission from 2014 until 2016 (*33*). Several features made the ZE outbreak suitable for this case study. First, the duration of human-to-human transmission was well-documented and followed a single cross-species transmission event between two reservoir host classes included in our models (Pterobat and primate). Second, genome sequences with known sampling dates were available from throughout the outbreak. Third, our models were able to detect primate signal in other Filovirus genomes (Figure 3A). These observations suggested that this combination of dataset and model should be suitable to specifically address whether short-term transmission in a novel host (~2 years) alters genomic biases and whether our models' predictions were altered by such short-term evolution.

We calculated genomic biases from 969 ZE genomes from Guinea, Liberia and Sierra Leone that had known months of collection. Two hundred GBMs were trained using the genomic features used in our main analysis, using iterative splitting of the data into training, optimization and validation sets as above. We removed the single historical representative of ZE to avoid including the same virus species in the training and prediction sets. The reservoir hosts of the genomes of the ZE viruses from the West African epidemic were predicted from trained models using the function *h2o.predict* in R. We recorded the BPS for each of the 11 reservoir host groups for each ZE sequence using the top 25% of models, scored according to validation set accuracy. Statistical analyses tested whether the BPS of the primate and Pterobat classes changed during the course of the epidemic using a GLMM with time (year and month represented by decimal, i.e., January 2015 = 2015.08) as a fixed effect and a random intercept for country of viral origin. We also tested whether each of the top 50 genomic biases changed during the epidemic using GLMMs with *p*-values corrected using the false discovery rate method. Term significance was estimated by comparing full models to models with only the random effect of country using an ANOVA (*34*).

Supplementary Text

Post-hoc analysis of misclassifications

To maximize the utility of our models for guiding future research and responses to emerging viruses, it was important to understand (*i*) whether model predictions contain signals of being erroneous or correct and (*ii*) how far off model predictions were when the top prediction was a misclassification. Using the bagged predictions from validation sets, we first evaluated whether stronger predictions (i.e., higher BPS for the top class) tended to be correct more often than equivocal predictions (i.e., the top class had only slightly higher BPS than other classes). For each prediction type, we fit logistic regressions relating the BPS of the top predicted class to the binary outcome of that prediction (correct or incorrect). All three prediction types showed significant positive relationships which implied that stronger predictions were likely to be true (Figure S13A-C). Indeed, strong predictions of arthropod-borne status (BPS>0.80) and vector identity (BPS>0.5) were nearly always correct. In contrast, incorrect reservoir predictions sometimes had strong BPS, indicating the greater challenge of predicting reservoirs (Figure S13A-C). Nevertheless, correct predictions spanned all prediction probabilities, and for reservoir host models and vector taxa models, even the weakest predictions outperformed accuracies expected by chance (number of classes^{-1}). For models predicting whether viruses have arthropod-borne transmission, a high baseline accuracy by chance alone (50%) meant that only GBM prediction strengths >0.65 outperformed chance. These results enhanced the interpretability of

orphan virus predictions by providing confidence estimates that predictions were correct and by showing that our models were less likely to project strong support to false predictions than to true predictions. While lower predicted probabilities may still be informative, these should be treated with greater caution and secondary predictions should be considered plausible (see below).

We next evaluated whether erroneous model predictions still contained useful information. This would be the case if misclassifications were not systematically biased by reservoir or vector taxonomy and the true host was instead ranked highly among alternative candidates. Heatmaps (Figure 2) showed minor biases between artiodactyls and primates, between Neoaves and Galloanserae birds and between Pterobats and Vespbats. To enhance these patterns, we replotted matrices following removal of the majority (correctly predicted diagonal) class (Figure S13F,G). Of the 123 viruses whose reservoir host was misclassified by GBMs, the true reservoir was most often the second rank predicted host (Figure S13D), such that the true reservoir occurred in the top 2 predictions for 81% of viruses and in the top 3 predictions for 86% of viruses (Figure 2C). For 12 viruses with misclassified vector taxa, the true vector was the second ranked prediction for 66.7% (Figure S13E), such that the true vector occurred within the top 2 predicted vectors for 95.9% of arthropod-borne viruses (Figure 2G). We omitted the ranking analysis of arthropod-borne transmission status because the binary nature of those models ensured that the second ranked prediction was correct.

Of the 527 viruses in the arthropod-borne prediction dataset, 5 (0.09%) consistently disagreed with the accepted transmission route across all model types (PN alone, PN and genomic features, genomic features alone). Chaoyang virus is a putative insect-specific Flavivirus that our models predicted to be arthropod-borne (BPS = 0.7), meaning it may also have a vertebrate host. Crimean-Congo hemorrhagic fever virus and Dugbe virus are Bunyaviruses that are believed to be transmitted predominately by ticks, but our models suggest an underappreciated role for direct transmission (BPS = 0.94 and 0.79, respectively). Moussa virus is Rhabdovirus detected only in arthropods but was predicted not to be arthropod-borne (BPS = 0.86), suggesting it may be insect-specific. Finally, Paraiso Escondido virus is a Flavivirus predicted to arthropod-borne (BPS = 0.73) despite suggestions that it is insect-specific based on failure to replicate vertebrate cells. Our reservoir prediction models also suggest an insect reservoir (BPS = 0.69). Additional work is required to determine the reasons for this apparent misclassification. BPS values reported above were from models with the highest family-specific bagged accuracy (Figure S13J).

Evolution of genomic biases during the 2014-2016 Zaire ebolavirus epidemic

GLMMs showed that most (41/50) genomic features changed during the course of the epidemic (false discovery corrected p-value < 0.08). While we observed significant decreases in 19 biases and significant increases in 22 biases, the magnitude of changes was small (Figure S17A). Intriguingly CpG dinucleotides declined across all codon positions (β = -0.002, marginal r^2 = 0.07, conditional r^2 = 0.31), at non-bridge codon positions (β = -0.003, marginal r^2 = 0.08, conditional r^2 = 0.3) and at bridge codon positions (β = -0.001, marginal r^2 = 0.02, conditional r^2 = 0.1). CpGs are widely suppressed in viruses of vertebrates, perhaps to evade immune defenses that target this sequence motif (*8*, *9*, *35*, *36*). Our results therefore suggest that this selection pressure intensified during human-to-human transmission of ZE, which is more broadly consistent with the idea that fine-tuning of genomic features to optimal levels for different host environments may play a role in viral adaptation.

Primate BPS of ZE varied between 0.5 and 3.6% and increased significantly during human-to-human transmission, though the effect size was small and explained relatively little variance (Figure S17B, $\beta = 0.002$, marginal $r^2 = 0.03$). Support for the Pterobat reservoir declined significantly during the same time period, though again explaining a small fraction of variance (Figure S17B, $\beta = -0.02$, marginal $r^2 = 0.06$).

Taken together, these results suggest that the genomic biases of ZE evolved directionally in the novel primate host, but that the magnitude of change was insufficient to obscure model predictions of reservoirs over a 2-year period of human-to-human transmission. The timescale at which transmission in novel hosts could be detected is likely to vary as a function of the host groups involved, viral mutation rates and generation times and other constraints on viral genomes, but will likely occur on the order of decades, rather than years. Additional analyses of viruses with sustained transmission in novel host groups may help clarify these timescales and their degree of consistency across viral groups.

Mixed host signals in the Middle East Respiratory Syndrome coronavirus genome

Middle East Respiratory Syndrome coronavirus (MERS) emerged as a novel human virus in 2012. Bats (specifically Vespbats, by our terminology) were initially posited as the most likely reservoir based on the discovery of closely related or identical coronaviruses in bats (*37*) and receptor usage patterns (*38*, *39*). However, serological evidence of infection in dromedary camels since the 1980s (*40*) and evidence for camel-to-human transmission (*41*) led to the emerging consensus that camels (an Artiodactyl) are the main source of human infections. Whether camels are now the sole reservoir (i.e., independently responsible for the long-term perpetuation of MERS in nature) and, if so, how long this has been the case remains poorly understood (*42*, *43*). Here, we provide a more detailed discussion of how our models deal with MERS, and potentially other viruses with similar epidemiological histories, to illustrate the limitations of our models, but also the safeguards that are built in to avoid overconfidence in singular erroneous predictions.

GBMs trained on PNs alone supported a bat reservoir of MERS, but with low confidence in distinguishing the two bat classes ($BPS_{Pterobat} = 0.18 \pm 0.13$ standard deviation; $BPS_{Vespbat} = 0.26 \pm 0.17$). Support for an Artiodactyl reservoir was low and similar to the remaining 8 reservoir groups ($0.05<BPS<0.08$). Adding genomic features to PN increased support for the Pterobat and Vespbat classes ($BPS_{Vespbat} = 0.55 \pm 0.37$; $BPS_{Pterobat} = 0.34 \pm 0.39$), had little influence on the probability of an Artiodactyl reservoir ($BPS_{Artiodactyl} = 0.06 \pm 0.14$) and reduced support for all other reservoir classes ($BPS < 0.01$). Finally, GBMs using only genomic features shifted support away from Pterobat ($BPS = 0.07 \pm 0.13$) to Artiodactyl ($BPS = 0.21 \pm 0.28$, a 3.7-fold increase in Artiodactyl support from the combined model) and retained support for a Vespbat reservoir ($BPS = 0.67 \pm 0.37$). These results imply that the PN contributes Pterobat signal which is inconsistent with genomic features, while genomic biases support Vespbats and Artiodactyls as candidate reservoirs, the two classes currently considered epidemiologically relevant. Thus, despite the fact that our models were designed to predict a single reservoir class, rather than host range, they may in some cases be able to detect mixed reservoir signals in viral genomes, though this requires confirmation in future studies.

MERS also illustrates the value of *post hoc* analysis of BPS to guide model interpretation. The low post-hoc probability of top prediction being true in the combined model (0.51 ± 0.037; estimated from the logistic regression in Figure S13A) and the expected gain in

accuracy from considering secondary predictions as plausible (Figure 2C, Figure S13D) implies that hosts beyond the top prediction could be considered. In the case of MERS, this would point to an Artiodactyl reservoir in the genomic feature only model or a Pterobat reservoir in the combined model. Generally, for making real-world decisions on research, surveillance and management, we recommend considering the predictions from both the combined and genomic feature models and viewing these predictions in light of what is known about the virus in question from other sources. When divergent predictions arise as a result of conflicting information between phylogenetic and genomic bias features as they did for MERS, we suggest greater confidence in the genomic bias models, since these had higher accuracy than the PN only models (Figure 2C, Figure S10B) and have a better chance of capturing the host associations of viruses that historically jumped between divergent host groups since, unlike PN, they do not depend entirely on evolutionary history (Figure 1B,C).

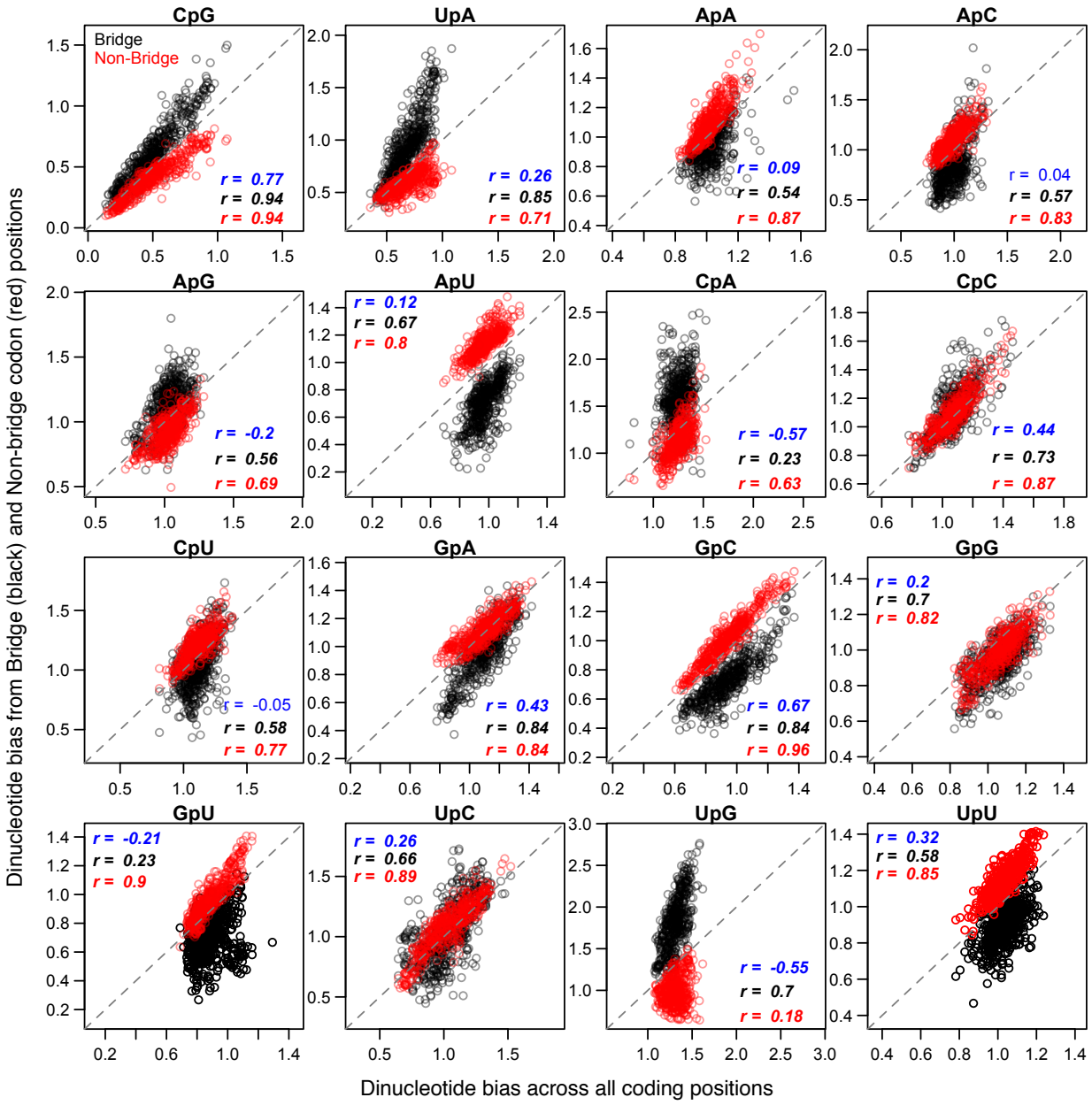

Fig. S1. Relationships between viral dinucleotide biases calculated from different codon positions. Panels shows the correlation between dinucleotide bias across all coding nucleotides in coding regions (x-axis) and the bias of dinucleotides that bridge adjacent codons (black) and those that do not ("non-bridge", red). The dashed grey line indicates the expectation that bridge and non-bridge dinucleotide biases perfectly match the pattern across all sites. Values in the bottom right are Pearson's correlation coefficients. Blue text indicates the correlation coefficient between bridge and non-bridge dinucleotides. Values in bold are statistically significant, including negative correlations. Each point represents a different virus species or strain.

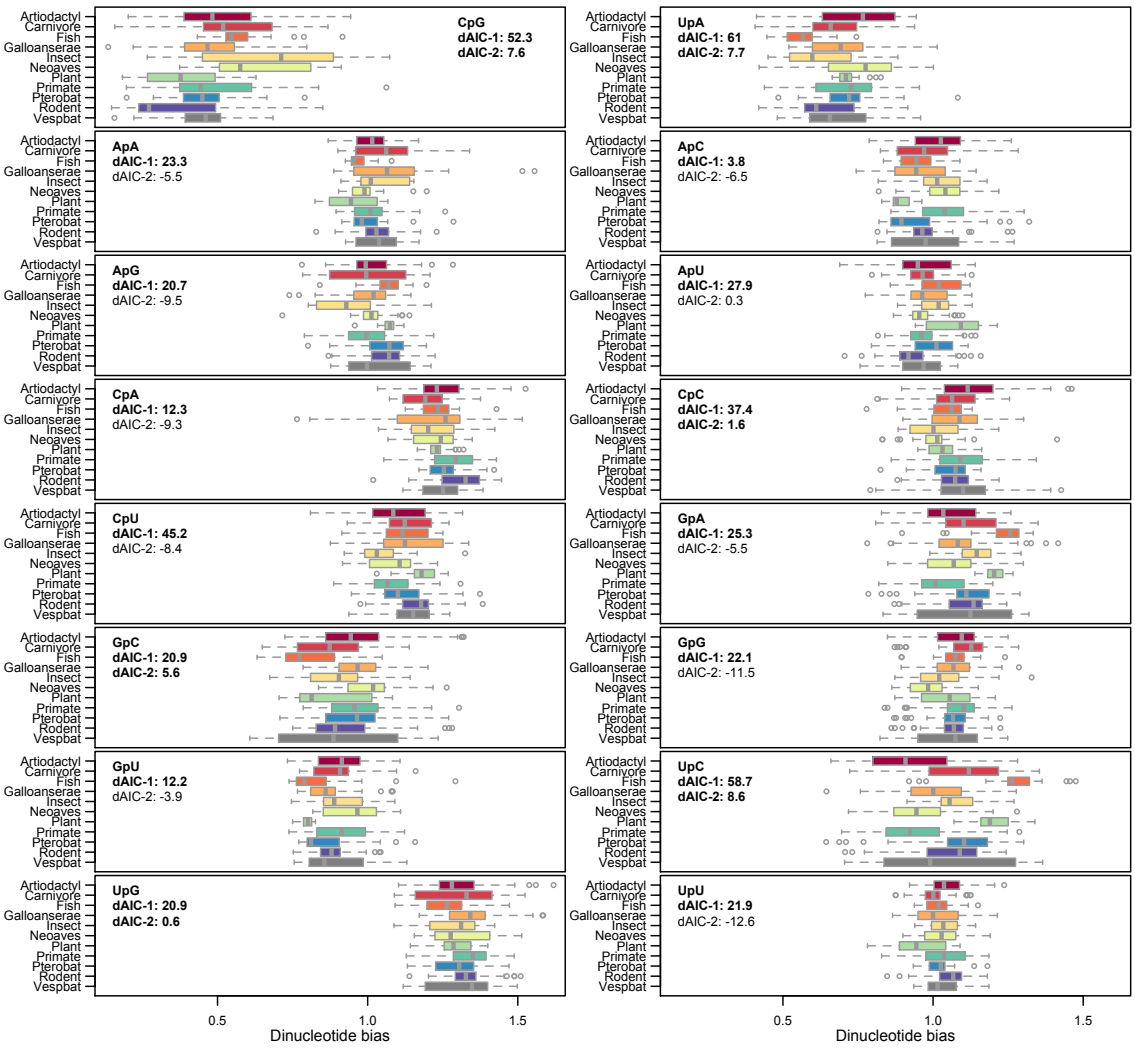

Fig. S2. Effects of reservoir host associations on viral dinucleotide bias. Boxplots show the distributions of dinucleotide frequencies across relatively common reservoir host taxa (N ≥ 15 viruses). Statistical support was evaluated as the increase in the Akaike information criterion (AIC) comparing generalized linear mixed models with and without a host group effect while including a random effect of viral group (dAIC-1). A second set of model comparisons tested cross-group effects of reservoir hosts on each dinucleotide by comparing models with and without a host effect while including a nested random of host within viral group (dAIC-2). Positive dAIC indicate increases in AIC after removing the reservoir host effect (i.e., poorer model fit). Bold dAIC values indicate statistical significance of the reservoir host effect according to F tests after Benjamini–Hochberg correction with a 10% false discovery rate. *P* values were considered statistically significant only if $p \leq 0.008$ (dAIC-1) and $p \leq 0.027$ (dAIC-2).

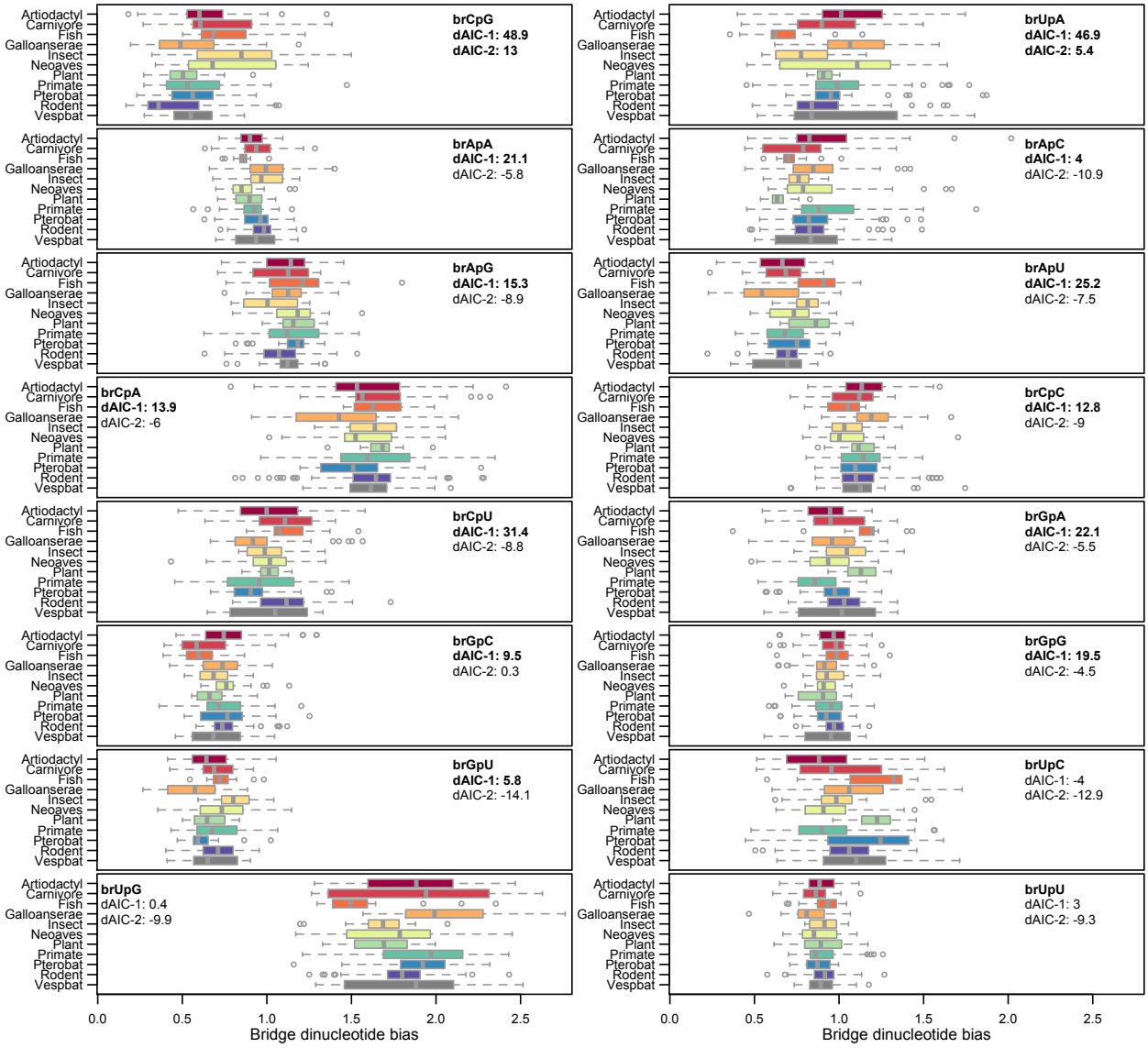

Fig. S3. Effects of reservoir host associations on viral dinucleotide bias at codon bridge positions. Boxplots show the distributions of dinucleotide frequencies calculated only at codon bridge positions across relatively common reservoir host taxa ($N \geq 15$ viruses). Statistical support was evaluated as the increase in the Akaike information criterion comparing generalized linear mixed models with and without a host group effect while including a random effect of viral group (dAIC-1). A second set of model comparisons tested cross-group effects of reservoir hosts on each dinucleotide by comparing models with and without a host effect while including a nested random of host within viral group (dAIC-2). Positive dAIC indicate increases in AIC after removing the reservoir host effect (i.e., poorer model fit). Bold dAIC values indicate statistical significance of the reservoir host effect according to F tests after Benjamini–Hochberg correction with a 10% false discovery rate. *P*-values were considered statistically significant only if $p \leq 0.008$ (dAIC-1) and $p \leq 0.005$ (dAIC-2).

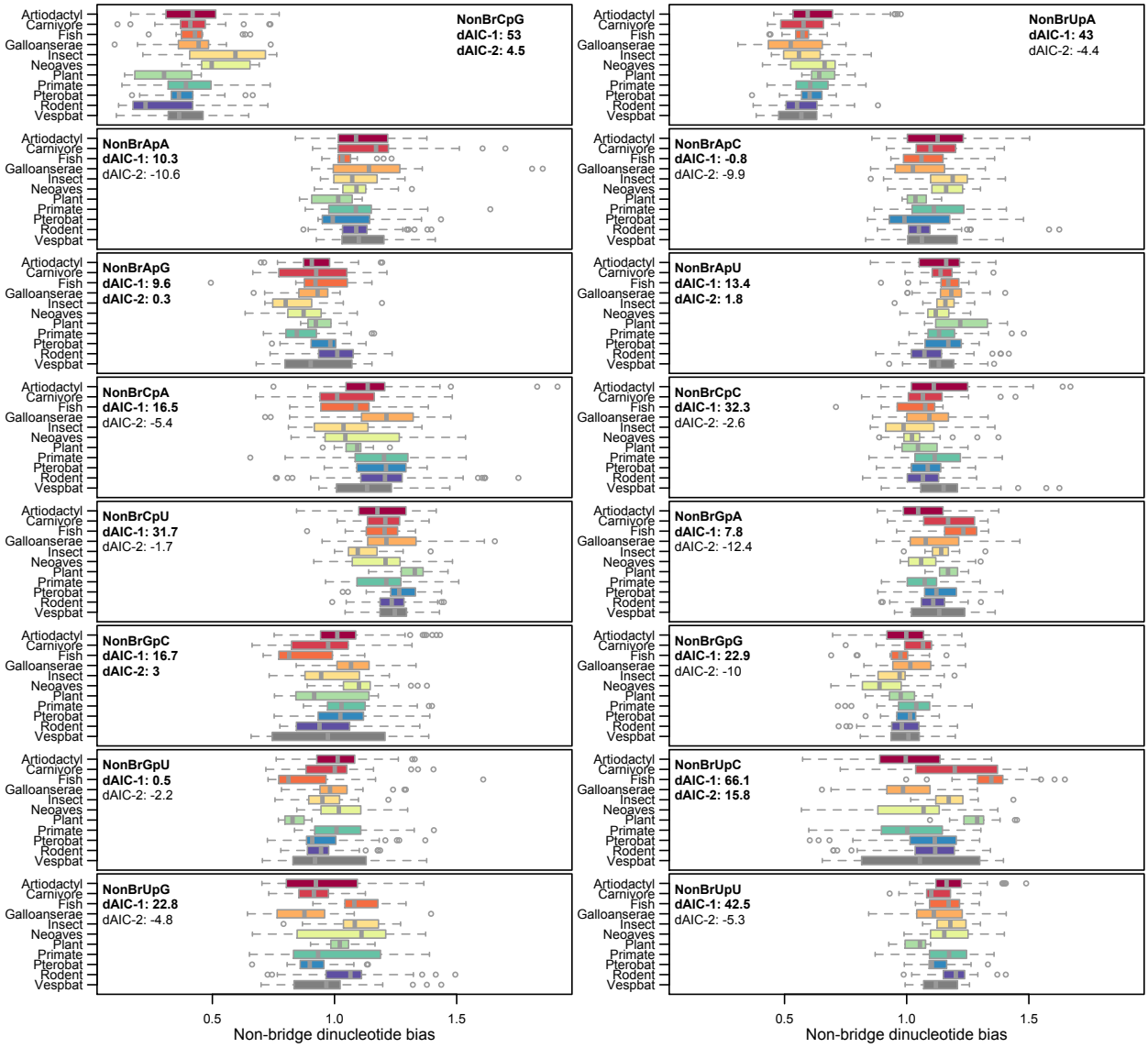

Fig. S4. Effects of reservoir host associations on viral dinucleotide bias at non-bridge codon positions. Boxplots show the distributions of dinucleotide frequencies calculated only at non-bridge codon positions across relatively common reservoir host taxa (N ≥ 15 viruses). Statistical support was evaluated as the increase in the Akaike information criterion comparing generalized linear mixed models with and without a host group effect while including a random effect of viral group (dAIC-1). A second set of model comparisons tested cross-group effects of reservoir hosts on each dinucleotide by comparing models with and without a host effect while including a nested random of host within viral group (dAIC-2). Positive dAIC indicate increases in AIC after removing the reservoir host effect (i.e., poorer model fit). Bold dAIC values indicate statistical significance of the reservoir host effect according to F tests after Benjamini–Hochberg correction with a 10% false discovery rate. *P*-values were considered statistically significant only if $p \leq 0.037$ (dAIC-1) and $p \leq 0.022$ (dAIC-2).

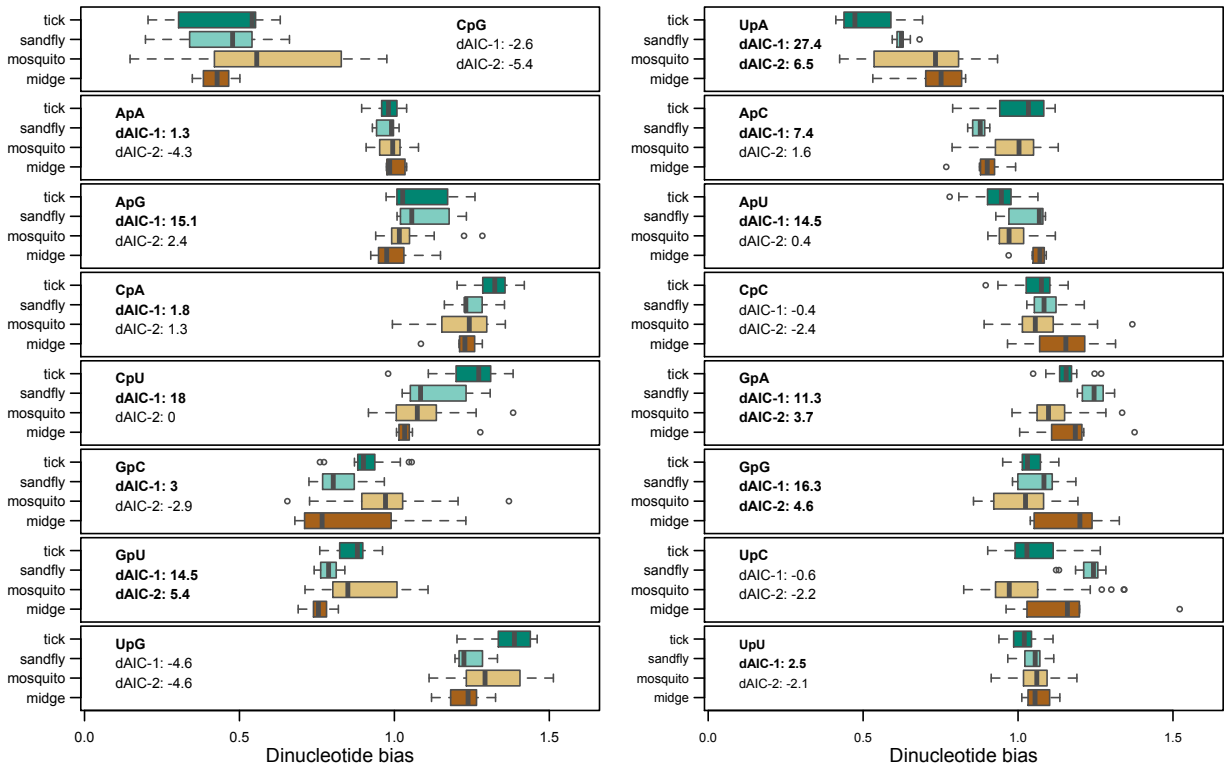

Fig. S5. Effects of arthropod vector associations on viral dinucleotide bias. Boxplots show the distribution of dinucleotide frequencies across relatively common arthropod vectors (N ≥ 8 viruses) of vertebrate viruses. Statistical support was evaluated as the increase in the Akaike information criterion comparing generalized linear mixed models with and without a vector group effect while including a random effect of viral group (dAIC-1). A second set of model comparisons tested cross-group effects of vectors on each dinucleotide by comparing models with and without a vector effect while including a nested random of vector within viral group (dAIC-2). Positive dAIC indicate increases in AIC after removing the vector effect (i.e., poorer model fit). Bold dAIC values indicate statistical significance of the reservoir host effect according to F tests after Benjamini–Hochberg correction with a 10% false discovery rate. *P*-values were considered statistically significant only if $p \leq 0.064$ (dAIC-1) and $p \leq 0.021$ (dAIC-2).

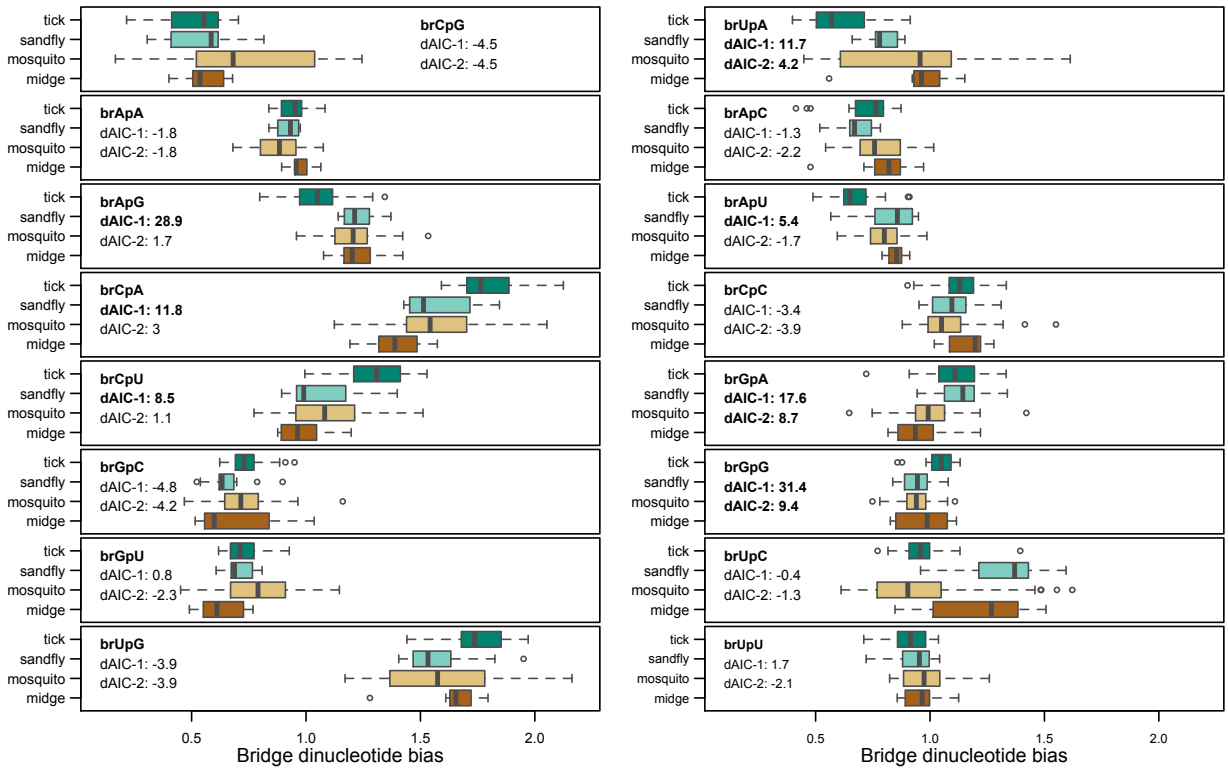

Fig. S6. Effects of arthropod vector associations on viral dinucleotide bias at codon bridge positions. Boxplots show the distribution of dinucleotide frequencies calculated only at codon bridge positions across relatively common arthropod vectors (N ≥ 8 viruses) of vertebrate viruses. Statistical support was evaluated as the increase in the Akaike information criterion comparing generalized linear mixed models with and without a vector group effect while including a random effect of viral group (dAIC-1). A second set of model comparisons tested cross-group effects of vectors on each dinucleotide by comparing models with and without a vector effect while including a nested random of vector within viral group (dAIC-2). Positive dAIC indicate increases in AIC after removing the vector effect (i.e., poorer model fit). Bold dAIC values indicate statistical significance of the reservoir host effect according to F tests after Benjamini–Hochberg correction with a 10% false discovery rate. *P*-values were considered statistically significant only if $p \leq 0.01$ (dAIC-1) and $p \leq 0.017$ (dAIC-2).

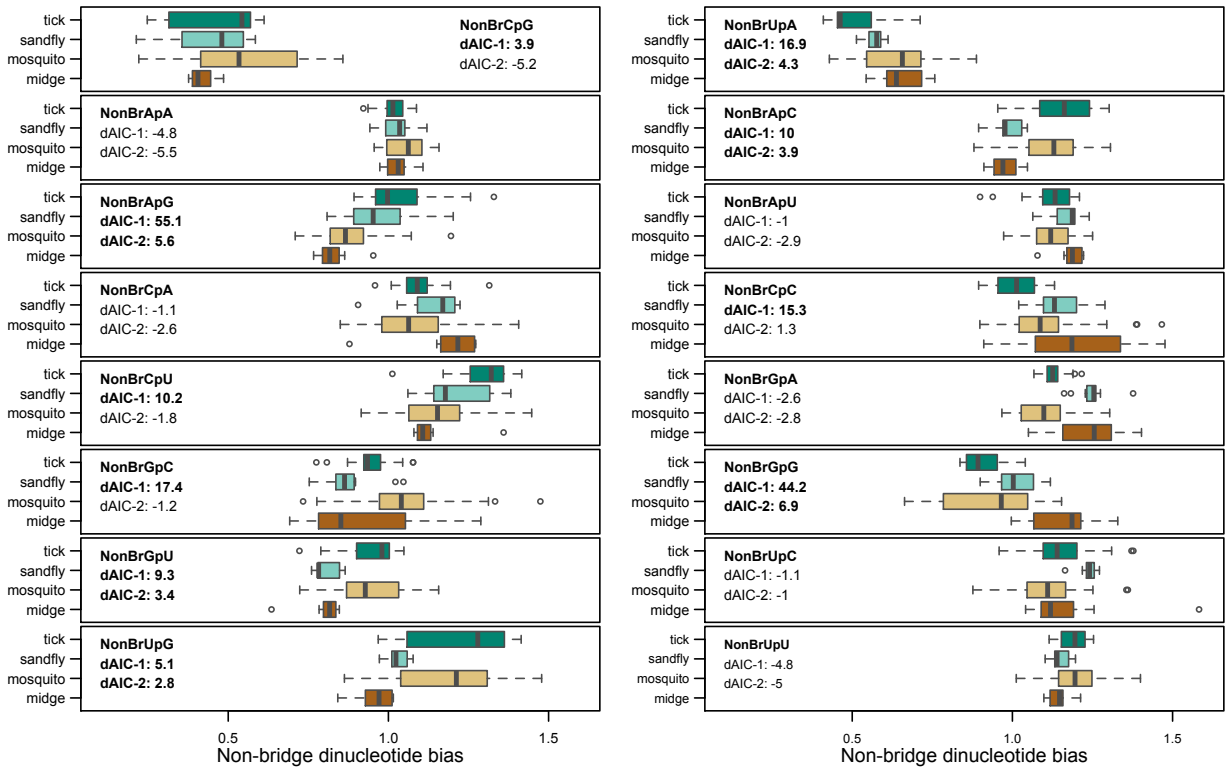

Fig. S7. Effects of arthropod vector associations on viral dinucleotide bias at non-bridge codon positions. Boxplots show the distribution of dinucleotide frequencies calculated only at non-bridge codon positions across relatively common arthropod vectors (N ≥ 8 viruses) of vertebrate viruses. Statistical support was evaluated as the increase in the Akaike information criterion comparing generalized linear mixed models with and without a vector group effect while including a random effect of viral group (dAIC-1). A second set of model comparisons tested cross-group effects of vectors on each dinucleotide by comparing models with and without a vector effect while including a nested random of vector within viral group (dAIC-2). Positive dAIC indicate increases in AIC after removing the vector effect (i.e., poorer model fit). Bold dAIC values indicate statistical significance of the reservoir host effect according to F tests after Benjamini–Hochberg correction with a 10% false discovery rate. *P*-values were considered statistically significant only if $p \leq 0.02$ (dAIC-1) and $p \leq 0.033$ (dAIC-2).

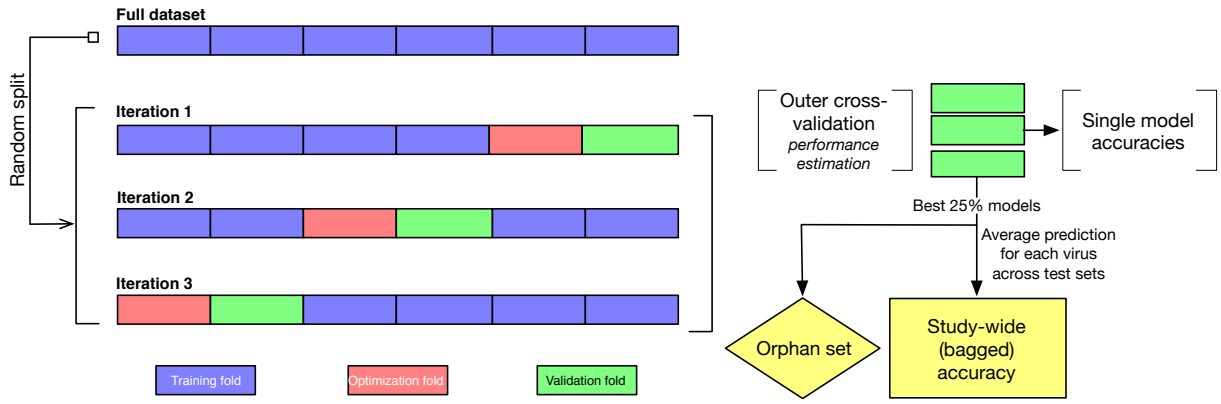

Fig. S8. Data splitting in machine learning analyses. At each round of model iterations, the full dataset was split into a training set (blue), optimization set (red) and validation set (green). For reservoir host prediction and arthropod-borne prediction, we trained each model on a randomly selected 70% of each class, optimized models on 15% of each class and estimated performance on the remaining 15% of each class, which were entirely withheld from training. Vector taxa prediction models were trained on 60% of each vector class. The remaining 40% of viruses were split 30/70 for optimization and validation. Following rounds of training, models were scored for overall accuracy (number correct/total) and per class accuracies. The best 25% of models based on overall accuracy were used for orphan predictions and to generate bagged predictions for each virus from models trained on different training sets.

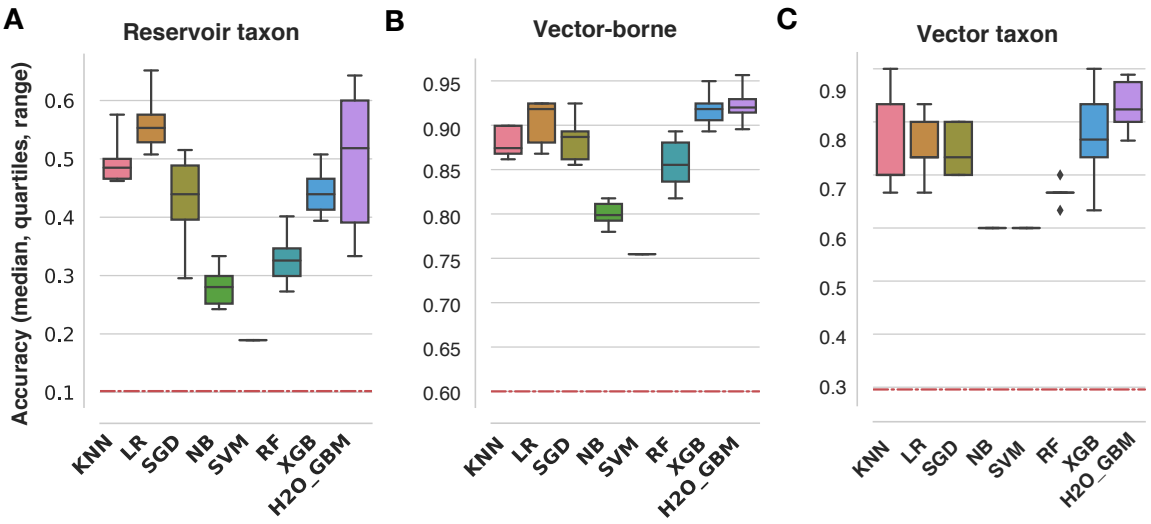

Fig. S9. Comparison of machine learning algorithms using all genomic features. Boxplots show the median, quartiles and range of accuracies for 8 algorithms from 10-fold cross validation of the same training set, each trained on the same genomic features for reservoir host prediction (A, 11 classes), arthropod-borne status (B, 2 classes), and vector taxon (C, 4 classes). The dashed red line (10.6%, 60%, and 29.6% respectively) indicates the expected accuracy from a null model where classes were assigned at random in proportion to their frequency in the training set.

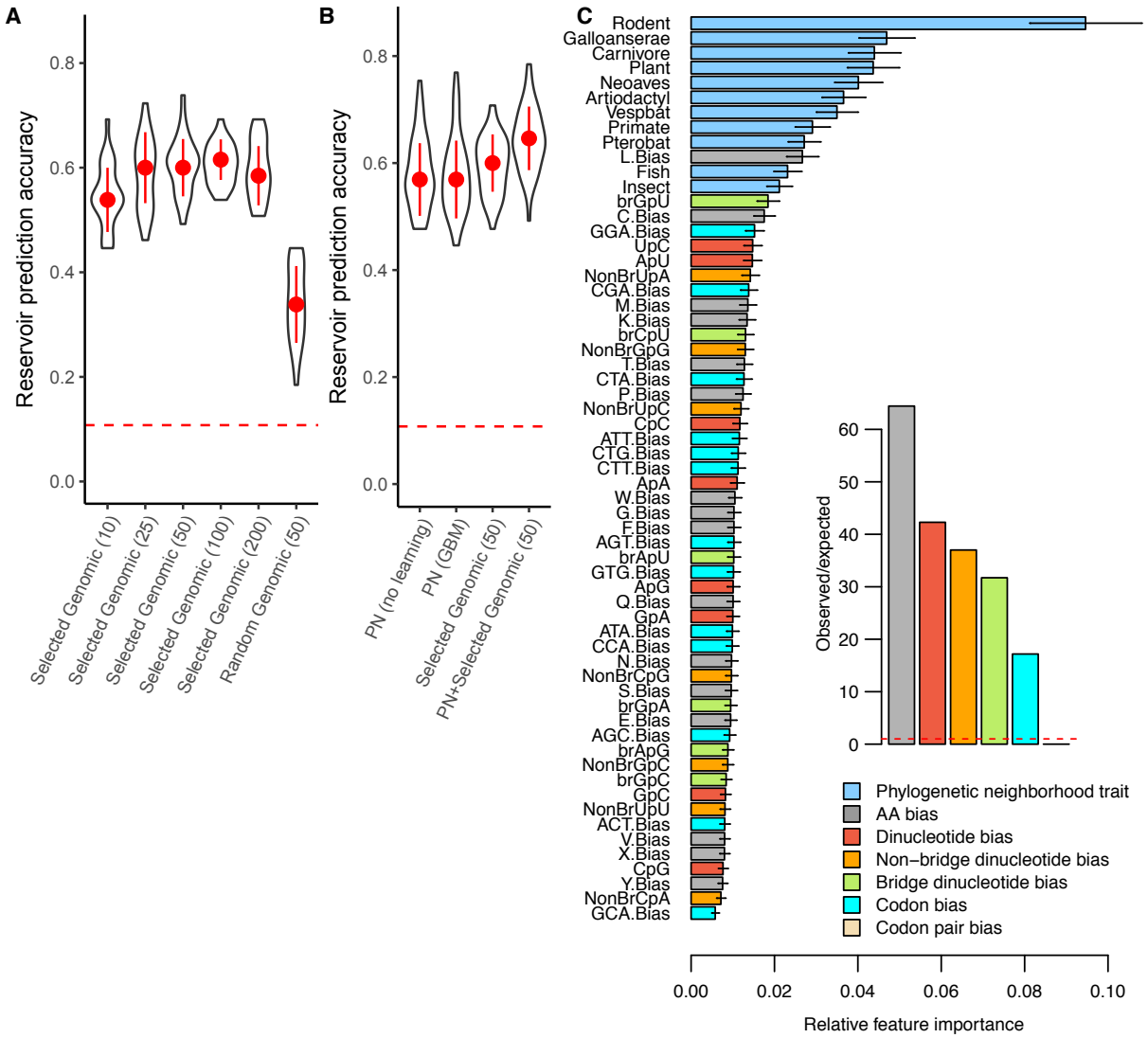

Fig. S10. Feature selection, model comparison and feature importance for reservoir host prediction. (A) Validation set accuracies across GBMs trained on different numbers of genomic features. Points and lines are medians and standard deviations from 25 models trained, optimized and validated on different data splits. The dashed line is the accuracy expected from a null model (classes randomly assigned in proportion to their frequency in the training set). The rightmost violin shows the accuracy from models trained on 50 randomly selected genomic features. Kruskal-Wallis tests with post-hoc Tukey comparisons indicated 50 as the minimal number of features that was significantly better than 10 (mean rank difference = 35.86, $p = 0.014$). We therefore retained 50 genomic features (47.3% of cumulative feature importance) in later analyses. (B) Accuracies for PNs without machine learning, and GBMs trained on PNs, 50 genomic biases and the combination of genomic biases and PNs (50 models each). (C) Feature importance averaged (± standard error) over 50 models. Genomic bias names use 3 letters for codons (e.g., ATG.Bias) and single letters for amino acids (e.g., L.Bias for Leucine bias). Longer strings are codon pair biases (e.g., ACC.T..CCG.P is bias in the use of AAC and CCG codons to encode sequential Threonine and Proline amino acids). The inset barplot shows the ratio between observed and expected frequencies of each class of genomic features among the top 50 features. The 1:1 line indicates features are as frequent in the top 50 as expected by chance.

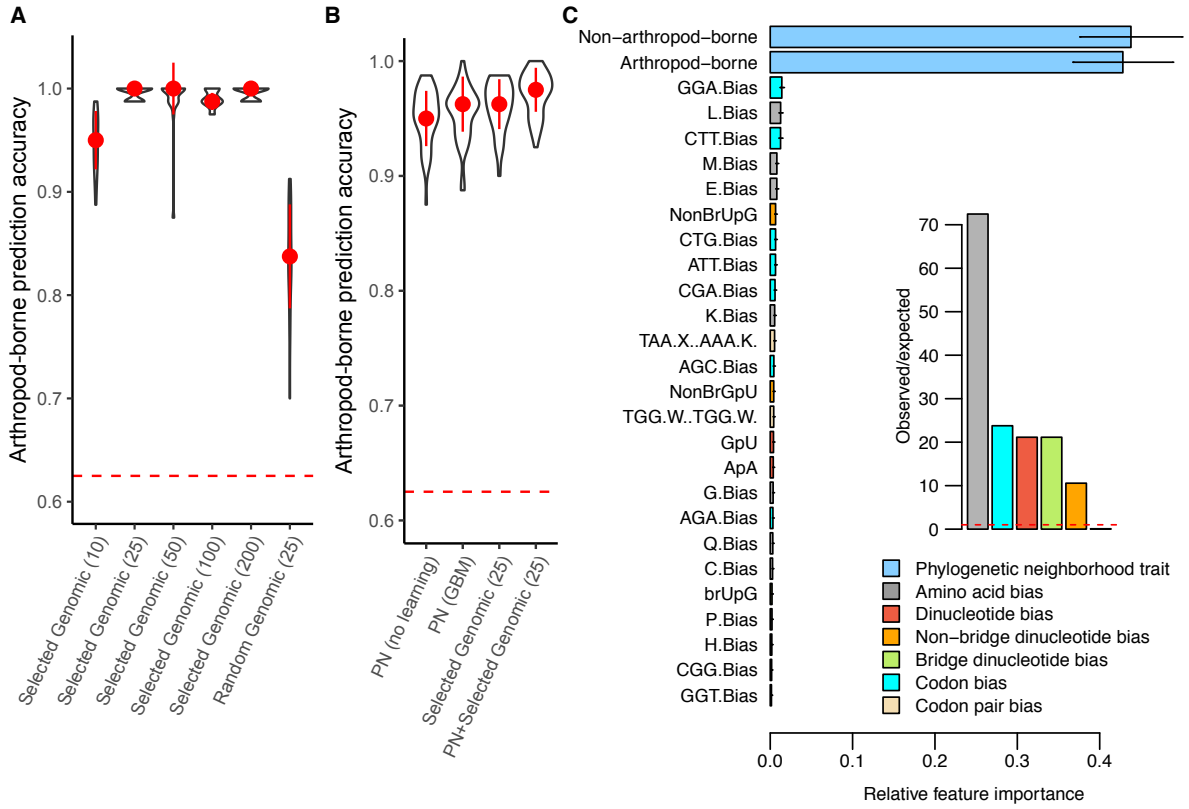

Fig. S11. Feature selection, model comparison and feature composition for prediction of whether viruses are transmitted via an arthropod vector. (A) Validation set accuracies across GBMs trained on different numbers of genomic features. Each violin summarizes 25 models trained and validated on different data splits. Models with >10 features were statistically equivalent, so we retained the top 25 genomic features for further analyses (71.8% of cumulative feature importance). (B) Validation set accuracies for PNs without machine learning, and GBMs trained on PNs alone, 25 genomic biases and the combination of 25 genomic biases and PNs (50 models each). (C) Feature importance (± standard error) of the combined model, averaged over 50 models and the distribution of the top 25 genomic features across categories of bias. Panel formatting follows Figure S10.

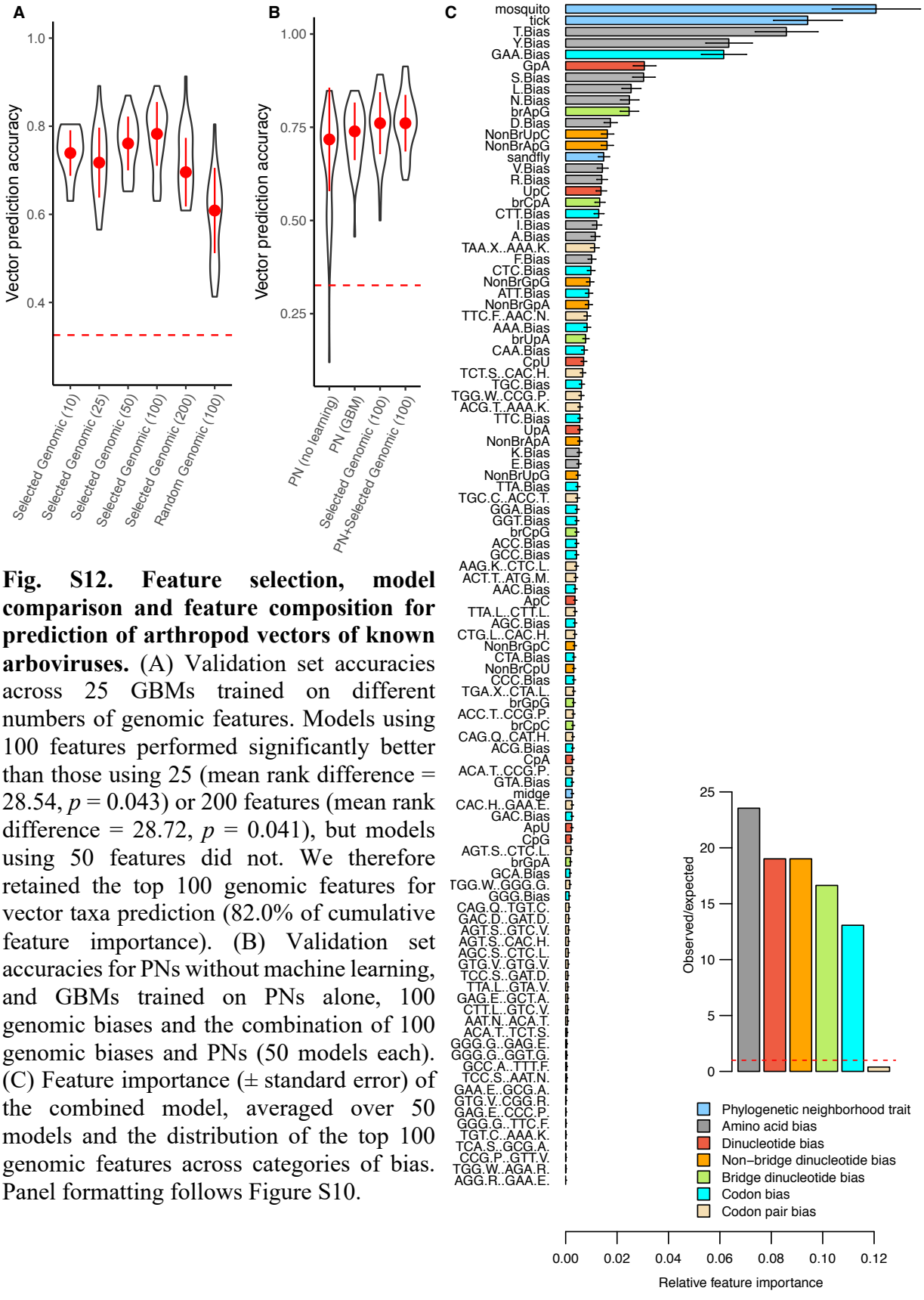

Fig. S12. Feature selection, model comparison and feature composition for prediction of arthropod vectors of known arboviruses. (A) Validation set accuracies across 25 GBMs trained on different numbers of genomic features. Models using 100 features performed significantly better than those using 25 (mean rank difference = 28.54, $p = 0.043$) or 200 features (mean rank difference = 28.72, $p = 0.041$), but models using 50 features did not. We therefore retained the top 100 genomic features for vector taxa prediction (82.0% of cumulative feature importance). (B) Validation set accuracies for PNs without machine learning, and GBMs trained on PNs alone, 100 genomic biases and the combination of 100 genomic biases and PNs (50 models each). (C) Feature importance (± standard error) of the combined model, averaged over 50 models and the distribution of the top 100 genomic features across categories of bias. Panel formatting follows Figure S10.

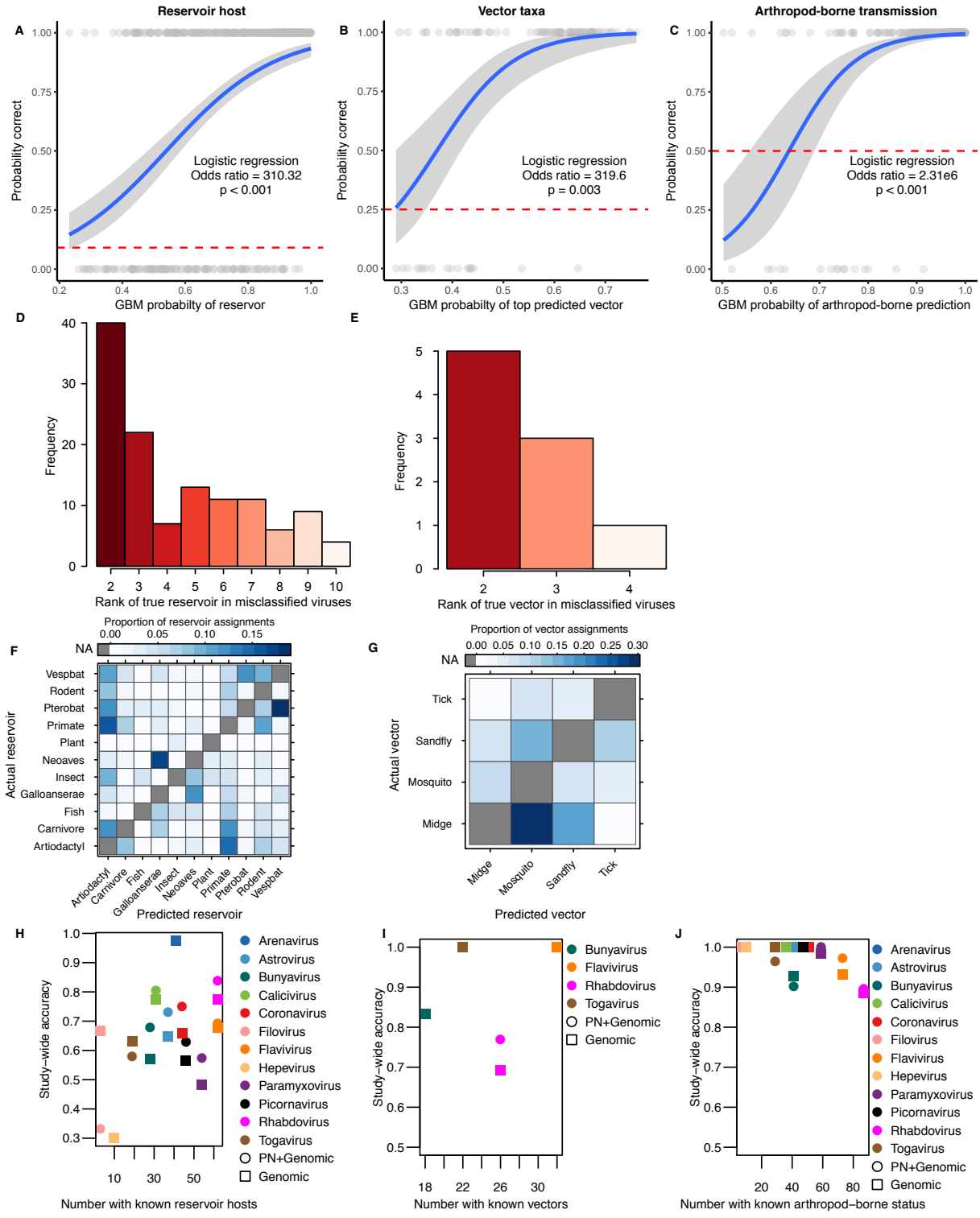

Fig. S13. *Post hoc* analyses of misclassification. (A-C) Logistic regression models for each prediction type, relating the strength of GBM predictions (BPS) to the prediction outcome. The blue line is the model prediction with standard error (grey shading). Points are observed outcomes (1 = correct, 0 = incorrect). The dashed line is the null accuracy, defined here as the number of classes^{-1}. (D-E) The rank of the true reservoir and vector for misclassified viruses.

Plots show the number of viruses within each category. (F-G) Heatmaps from the main text (Figure 2) with correct predictions omitted to highlight patterns in misclassification. (H-J) Scatterplots showing the effects of sample sizes in each viral group on accuracy for the genomic trait only model (squares) and for the combined genomic trait and PN model (circles). Reservoir hosts from larger viral groups were predicted more accurately in the combined model (Pearson's correlation: $r = 0.65$ for the combined model, $p = 0.02$), with a weaker non-significant trend in the genomic feature only model, $r = 0.31$, $p = 0.33$) for the genomic trait only model, $p < 0.05$). The size of viral groups was not related to vector prediction accuracy ($p > 0.05$) but was slightly negatively related to accuracy of predictions of arthropod-borne transmission (genomic trait model: $r = -0.70$, $p = 0.012$; combined model: $r = -0.46$, $p = 0.14$).

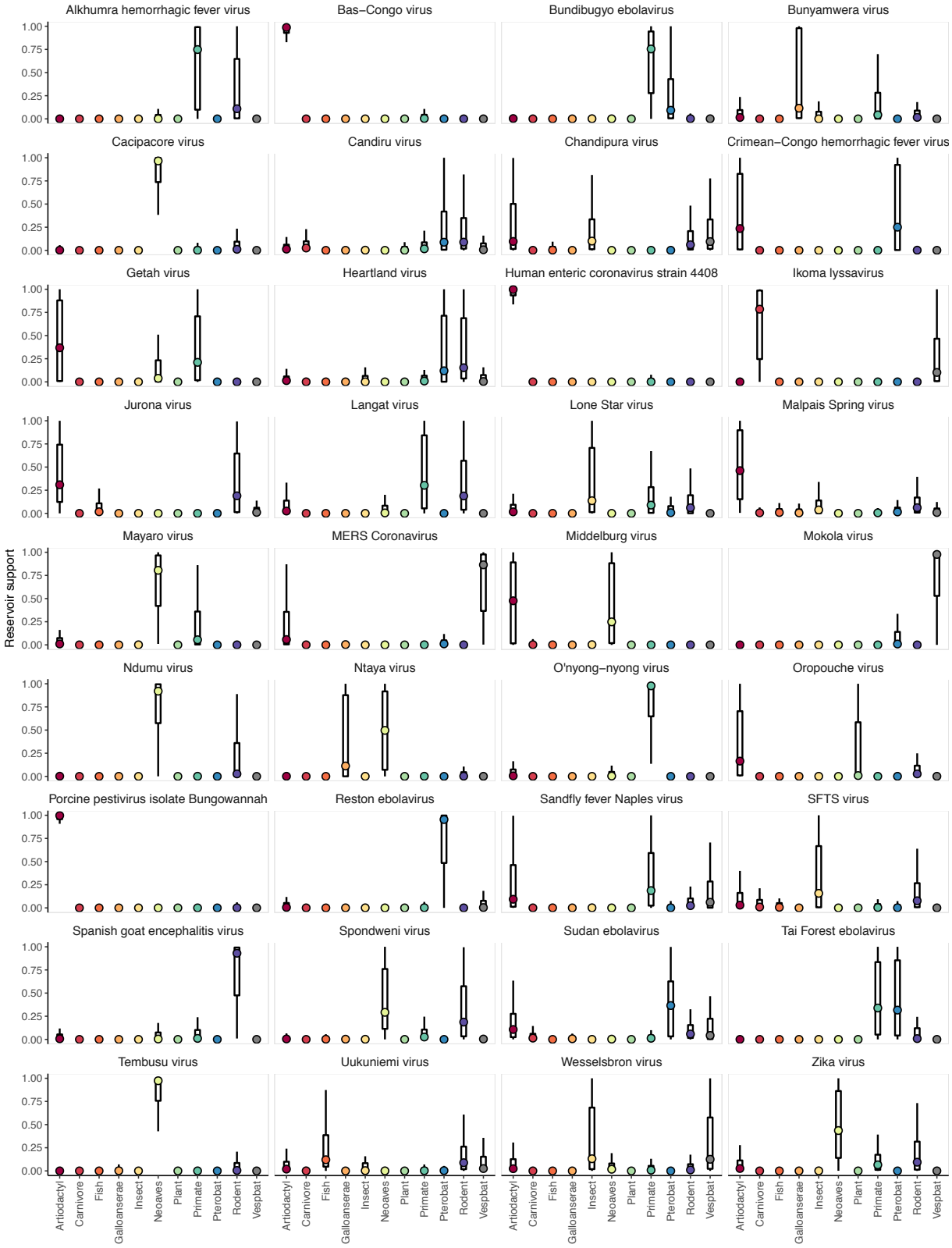

Fig. S14. Reservoir predictions for 36 emerging orphan viruses. Boxplots show the probability distribution of each reservoir group across the top 25% of GBMs. SFTS=Severe Fever with Thrombocytopenia Syndrome.

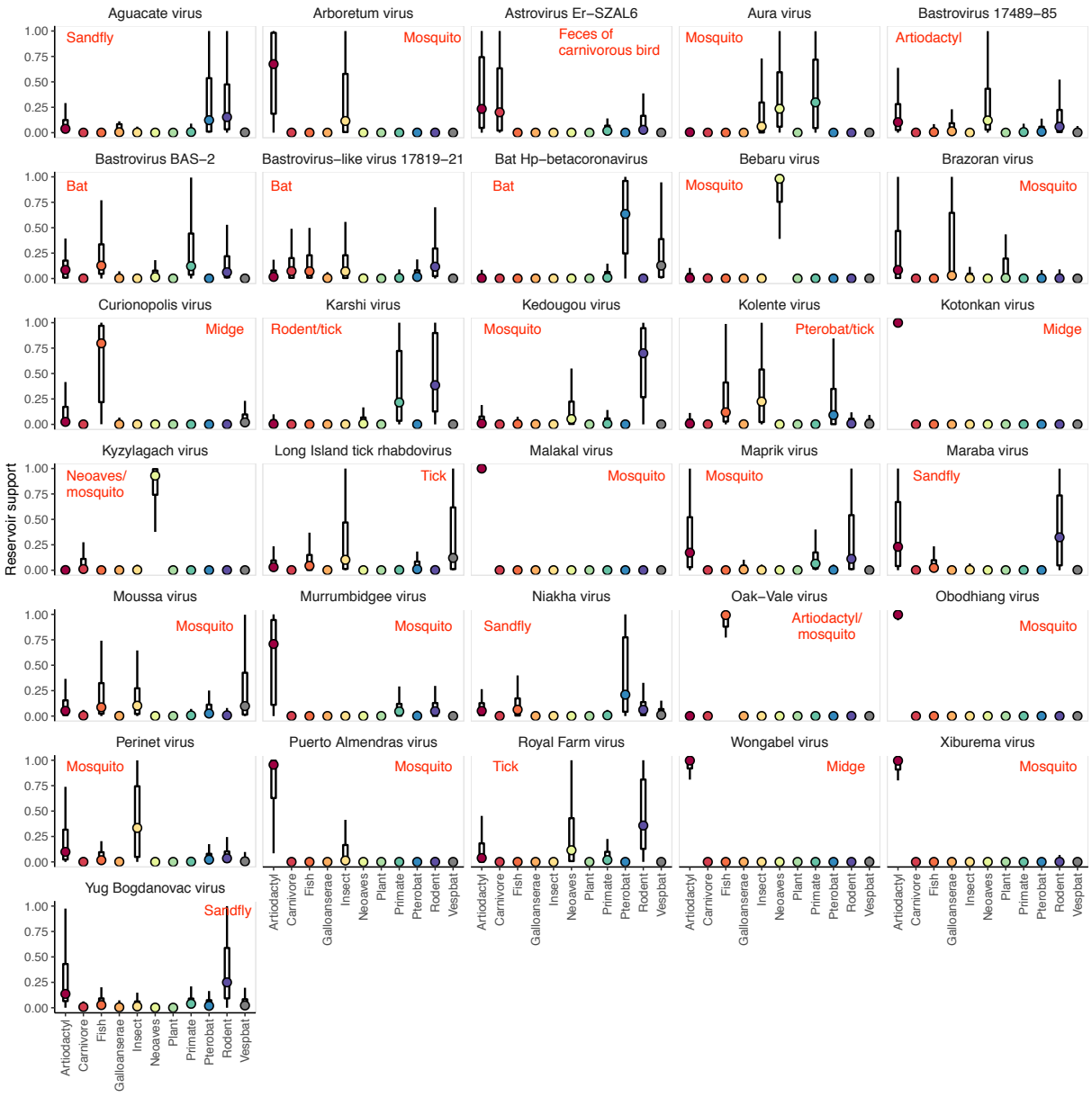

Fig. S15. Reservoir predictions for 31 orphan viruses detected through active surveillance of reservoirs or vectors. Boxplots show the probability distribution of each reservoir group across the top 25% of GBMs. Red text indicates the host where each virus was detected.

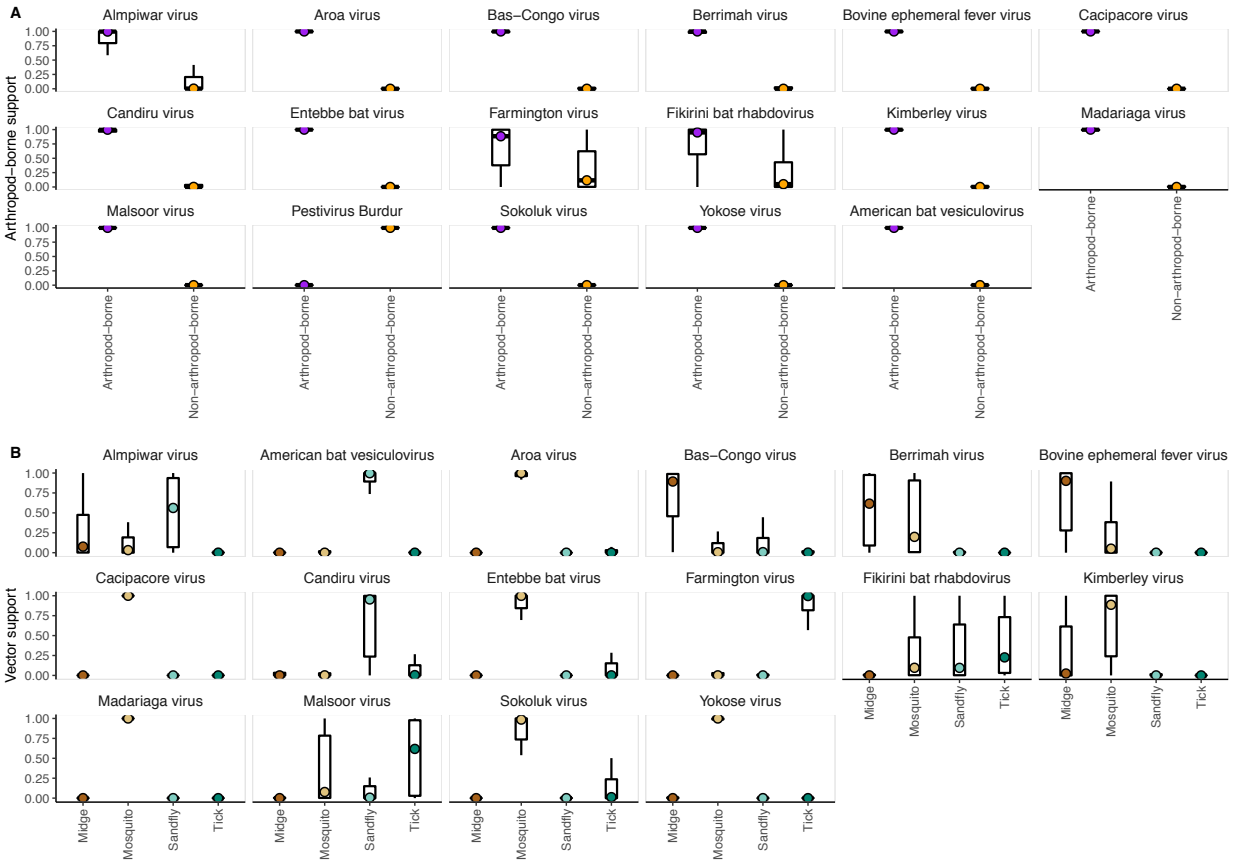

Fig. S16. Predictions of arthropod-borne transmission status and vector type for viruses with unknown transmission routes or vector taxa. Boxplots show the probability arthropod-borne transmission (A) and the probability of each vector group inferred from the top 25% GBMs.

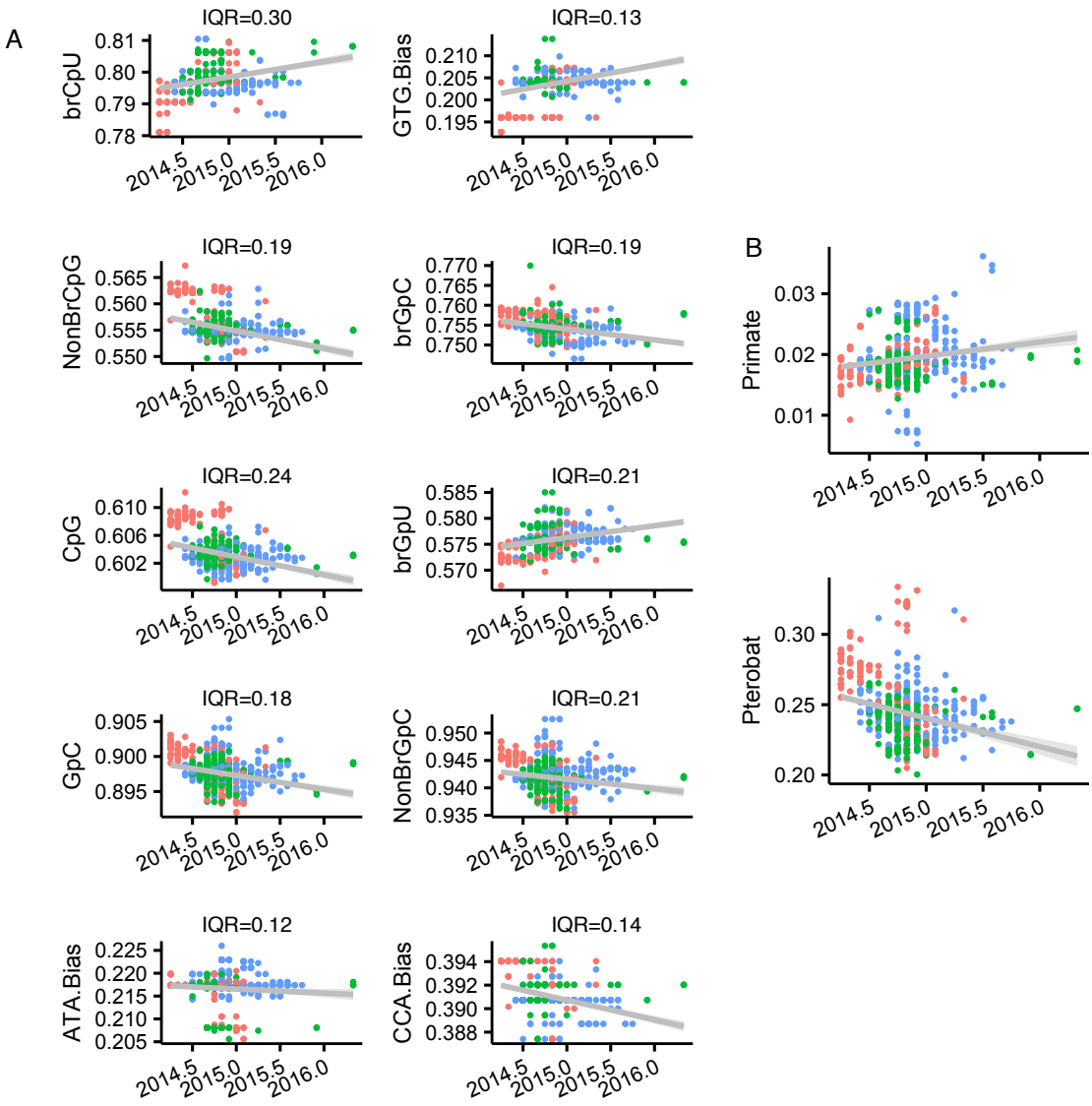

Fig. S17. Temporal dynamics of viral genomic biases and model predictions during the West African Zaire ebolavirus epidemic. (A) The 10 genomic features with the largest absolute effect sizes according to GLMMs. Points are colored according to country of origin (red = Guinea; green = Liberia; blue = Sierra Leone). IQR = interquartile range for all viruses in the reservoir host model. (B) Bagged prediction scores for the primate and Pterobat reservoir classes over the course of human-to-human transmission. Fitted lines are linear models with 95% confidence interval (shading) for illustrative purposes only All relationships shown were statistically significant in GLMMs

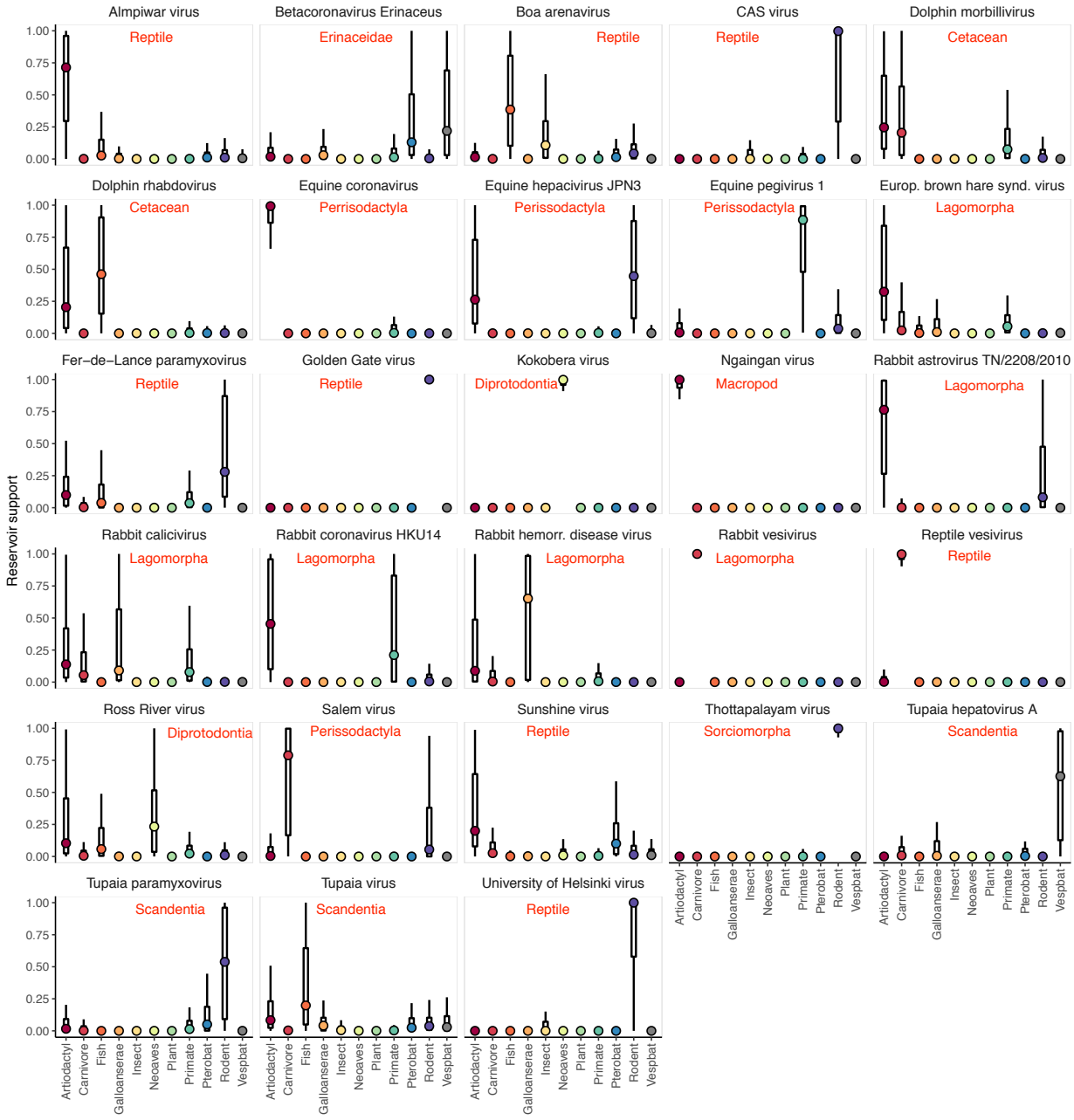

Fig. S18. Predictions of reservoir hosts for viruses from underrepresented host groups. Boxplots show the probability distribution of each reservoir group across the top 25% of GBMs.